Trypanosomal variant surface glycoprotein expression in human African trypanosomiasis patients

Jaime So^{1**}, Sarah Sudlow^{1**}, Abeer Sayeed¹, Tanner Grudda¹, Stijn Deborggraeve^{2#}, Dieudonné Mumba Ngoyi³, Didier Kashiama Desamber⁴, Bill Wickstead⁵, Veerle Lejon⁶, and Monica R. Mugnier^{1*}

¹Department of Molecular Microbiology and Immunology, Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland, United States of America

²Department of Biomedical Sciences, Institute of Tropical Medicine, Antwerp, Belgium

³Department of Parasitology, Institut National de Recherche Biomédicale, Kinshasa, Democratic Republic of the Congo

⁴Programme Nationale de Lutte contre la Trypanosomiase Humaine Africaine, (PNLTHA), Ministry of Health, Kinshasa, Democratic Republic of the Congo

⁵ School of Life Sciences, Queen's Medical Centre, University of Nottingham, Nottingham, NG7 2UH, United Kingdom

⁶ UMR-177 Intertryp, Institut de Recherche pour le Développement, Centre de Coopération Internationale en Recherche Agronomique pour le Développement, University of Montpellier, Montpellier, France

******These authors contributed equally to the work

Current address: Médecins Sans Frontières - Access Campaign, Geneva, Switzerland

***** Corresponding author

E-mail: mmugnie1@jhu.edu (MM)

Abstract

Trypanosoma brucei gambiense, an extracellular protozoan parasite, is the primary causative agent of human African trypanosomiasis. *T. b. gambiense* is endemic to West and Central Africa, where it is transmitted by the bite of infected tsetse flies. In the bloodstream of an infected host, the parasite evades antibody recognition by altering the Variant Surface Glycoprotein (VSG) that forms a dense coat on its cell surface through a process known as antigenic variation. Each VSG has a variable N-terminal domain exposed to the host and a less variable C-terminal domain that is partially hidden from host antibodies. Our lab developed VSG-seq, a targeted RNA-seq method, to study VSG gene expression in *T. brucei*. Studies using VSG-seq to characterize antigenic variation in a mouse model have revealed marked diversity in VSG expression within parasite populations, but this finding has not yet been validated in a natural human infection. Here, we used VSG-seq to analyze VSGs expressed in the blood of 12 patients infected with *T. b. gambiense*. The number of VSGs identified per patient ranged from one to 14; notably, two VSGs were shared by more than one patient. Analysis of expressed VSG N-terminal domain types revealed that 82% of expressed VSGs encoded a type B N-terminus, a bias not seen in datasets from other *T. brucei* subspecies. C-terminal types in *T. b. gambiense* infection were also restricted. These results demonstrate a bias either in the underlying VSG repertoire of *T. b. gambiense* or in this subspecies' expression of VSGs during infection. This work demonstrates the feasibility of using VSG-seq to study antigenic variation in human infections and highlights the importance of understanding VSG repertoires in the field.

Author Summary

Human African Trypanosomiasis is a neglected tropical disease primarily caused by the extracellular parasite *Trypanosoma brucei gambiense*. To avoid elimination by the host, these parasites repeatedly replace their dense surface coat of Variant Surface Glycoprotein (VSG). Despite the important role of VSGs in prolonging infection, VSG expression during natural human infections is poorly understood. A better understanding of natural VSG gene expression dynamics can clarify the mechanisms that *T. brucei* uses to alter its VSG coat and improve how trypanosomiasis is diagnosed in humans. We analyzed the expressed VSGs detected in the blood of patients with trypanosomiasis. Our findings indicate that parasites in both natural and experimental infections express diverse sets of VSGs.

Introduction

Human African Trypanosomiasis (HAT) is caused by the protozoan parasite *Trypanosoma brucei*. *T. brucei* and its vector, the tsetse fly, are endemic to sub-Saharan Africa [1]. There are two human-infective *T. brucei* subspecies: *T. b. gambiense*, which causes chronic infection in West and Central Africa (~98% of cases), and *T. b. rhodesiense*, which causes acute infection in East and Southern Africa (~2% of cases) [2,3]. In humans, infections progress from an early stage, usually marked by a fever and body aches, to a late stage associated with severe neurological symptoms that begins when the parasite crosses the blood-brain barrier [4]. HAT is considered fatal without timely diagnosis and treatment. While around 50 million people are at risk of infection [5], the number of annual human infections has declined significantly in recent years, with only 864 cases reported in 2019 [6]. The World Health Organization is working towards zero human transmissions of HAT caused by *T. b. gambiense* (gHAT) by 2030 [7]. Current public health initiatives to control the disease depend on case detection and treatment, complemented with vector control.

Prospects for developing a vaccine are severely confounded by the ability of African trypanosomes to alter their surface antigens [8]. As *T. brucei* persists extracellularly in blood, lymph, and tissue fluids, it is constantly exposed to host antibodies [9–12]. The parasite periodically changes its dense Variant Surface Glycoprotein (VSG) coat to evade immune recognition. This process, called antigenic variation, relies on a vast collection of thousands of VSG-encoding genes [13–16]. *T. brucei* also continually expands the number of usable antigens by constructing mosaic VSGs through one or more recombination events between individual VSG genes [17,18].

The VSG contains two domains: a variable N-terminal domain of ~350-400 amino acids, and a less variable C-terminal domain of ~40-80 amino acids, characterized by one or two conserved groups of four disulfide-bonded cysteines [13,19]. On the surface of trypanosomes, the VSG N-terminal domain is readily exposed to the host. In contrast, the C-terminal domain is proximal to the plasma membrane and largely hidden from host antibodies [20–22]. The N-terminal domain is classified into two types, A and B, each further distinguished into subtypes (A1-3 and B1-2), while the C-terminal domain has been classified into six types (1-6) [13,19]. These classifications are based on protein sequence patterns anchored by the conservation of cysteine residues, but the biological implications of VSG domain types have not been investigated.

Little is known about how the large genomic repertoire of VSGs is used in natural infections; the number and diversity of VSGs expressed by wild parasite populations remain unknown. One VSG in particular, LiTat 1.3, has been identified as an antigen against which many gHAT patients have antibodies [23]. As a result, LiTat 1.3 is the main target antigen in the primary serological screening tool for gHAT: the card agglutination test for trypanosomiasis (CATT/*T. b. gambiense*) [24]. Despite the widespread use of the CATT to screen for gHAT, there are shortcomings. Most notably, it relies on immunity against a variable antigen, and some *T. b. gambiense* strains lack the LiTat 1.3 gene entirely [25,26]. More recently developed rapid diagnostic tests use a

combination of native LiTat1.3 and another VSG, LiTat1.5 [27,28], or the combination of a VSG with the invariant surface glycoprotein ISG 65 [29]. Currently, there is no serological test for diagnosis of infection with *T. b. rhodesiense*.

Given the role of VSGs during infection and their importance in gHAT screening tests, a better understanding of VSG expression dynamics could inform the development of improved screening tests while providing insight into the molecular mechanisms of antigenic variation. We developed VSG-seq, a targeted RNA-sequencing method that identifies the VSGs expressed in any population of *T. brucei* and measures the prevalence of each VSG in the population [30]. In a proof-of-principle study, we used VSG-seq to gain insight into the number and diversity of VSGs expressed during experimental mouse infections [30]. This proof-of-principle study revealed significant VSG diversity within parasite populations in each animal, with many more variants expressed at a time than the few thought to be sufficient for immune evasion. This diversity suggested that the parasite's genomic VSG repertoire might be insufficient to sustain a chronic infection, highlighting the potential importance of the recombination mechanisms that form new VSGs[13,17]. The study also showed that many of the same VSGs were expressed in different infections, supporting previous reports of a “semi-predictable” order to VSG switching in *T. brucei* culture and animal models of infection [17,31,32].

Recently, similar high-throughput sequencing methods have been used to characterize antigenic variation in experimental infections of natural hosts for two related African trypanosome species, *T. vivax* and *T. congolense* [33–35]. These studies suggest that the mechanism for antigen production in some animal parasites may be different from the *T. brucei* model. It remains unknown whether findings from studies of antigenic variation using experimental models translate to natural *T. brucei* infections. To our knowledge, only one study has investigated VSG expression in wild *T. brucei* isolates [36]. For technical reasons, this study relied on RNA isolated from parasites passaged through small animals after collection from the natural host. As VSG expression may change during passage, the data obtained from these samples is somewhat difficult to interpret. To better understand the characteristics of antigenic variation in natural *T. brucei* infections, we sought to analyze VSG expression in *T. brucei* field isolates from which RNA was directly extracted.

In the present study, we used VSG-seq to determine the number and diversity of VSGs expressed by *T. b. gambiense* in the blood of 12 patients with a *T. b. gambiense* infection. To complement these data, we also analyzed published datasets from both experimental mouse infections and *T. b. rhodesiense* patients. Our analysis revealed diverse expression of VSGs in natural *T. brucei* infections and distinct biases in VSG expression that may be unique to the *T. b. gambiense* subspecies.

Methods

Ethics statement

The blood specimens from *T.b. gambiense* infected patients were collected within the projects, “Longitudinal follow-up of CATT seropositive, trypanosome negative individuals (SeroSui)” and “An integrated approach for identification of genetic determinants for susceptibility for trypanosomiasis (TrypanoGEN)” [37]. In France, the SeroSui study received approval from the Comité Consultatif de Déontologie et d’Ethique (CCDE) of the French National Institute for Sustainable Development Research (IRD), May 2013 session. In Belgium, the study received approval from the Institutional Review Board of the Institute of Tropical Medicine (reference 886/13) and the Ethics Committee of the University of Antwerp (B300201318039). In the Democratic Republic of the Congo, the projects SeroSui and TrypanoGEN were approved by the Ministry of Health through the Ngaliema Clinic of Kinshasa (references 422/2013 and 424/2013). Participants gave their written informed consent to participate in the projects. For minors, additional written consent was obtained from their legal representative.

Patient enrollment and origin map

Patients originated from the Democratic Republic of the Congo (DR Congo) and were identified over six months in the second half of 2013. This identification occurred either during passive screening at the center for HAT diagnosis and treatment at the hospital of Masi Manimba, or during active screening by the mobile team of the national sleeping sickness control program (PNLTHA) in Masi Manimba and Mosango health zones (Kwilu province, DR Congo).

Individuals were screened for the presence of specific antibodies in whole blood with the CATT test. For those reacting blood positive in CATT, we also tested twofold serial plasma dilutions of 1/2-1/32 were also tested and determined the CATTCATT end titer was determined. CATTCATT positives underwent parasitological confirmation by direct microscopic examination of lymph (if enlarged lymph nodes were present), and examination of blood by the mini-anion exchange centrifugation technique on buffy coat [38]. Individuals in whom trypanosomes were observed underwent lumbar puncture. The cerebrospinal fluid was examined for white blood cell count and the presence of trypanosomes to determine the disease stage and select the appropriate treatment. Patients were questioned about their place of residence. The geographic coordinates of their corresponding villages were obtained from the Atlas of HAT [39] and plotted on a map of the DR Congo using ArcGIS® software by Esri. Distances were determined and a distance matrix generated (see Supplemental Table 2).

Patient blood sample collection and total RNA isolation

A 2.5 mL volume of blood was collected from each patient in a PAXgene Blood RNA Tube. The blood was mixed with the buffer in the tube, aliquoted in 2 mL volumes and frozen in liquid nitrogen for a maximum of two weeks. After arrival in Kinshasha, tubes

were stored at -70°C. Total RNA was extracted and isolated from each blood sample as previously described [40].

Estimation of parasitemia

Two approaches were used to estimate parasitemia. First, a 9 mL volume of blood on heparin was centrifuged, 500 microliters of the buffy coat were taken up and trypanosomes were isolated using the mini-anion exchange centrifugation technique. After centrifugation of the column eluate, the number of parasites visible in the tip of the collection tube were estimated. Second, Spliced Leader (SL) RNA expression levels were measured by real-time PCR as previously described [40]. A Ct value was determined for each patient blood sample. Real-time PCR was performed on RNA samples before reverse transcription to verify the absence of DNA contamination.

RNA sequencing

DNase I-treated RNA samples were cleaned up with 1.8x Mag-Bind TotalPure NGS Beads (Omega Bio-Tek, # M1378-01). cDNA was generated using the SuperScript III First-strand synthesis system (Invitrogen, 18080051) according to manufacturer's instructions. 8 microliters of each sample (between 36 and 944 ng) were used for cDNA synthesis, which was performed using the oligo-dT primer provided with the kit. This material was cleaned up with 1.8x Mag-Bind beads and used to generate three replicate library preparations for each sample. These technical replicates were generated to ensure that any VSGs detected were not the result of PCR artifacts[41,42].

Because we expected a low number of parasites in each sample, we used a nested PCR approach to prepare the VSG-seq libraries. First, we amplified *T. brucei* cDNA from the parasite/host cDNA pool by PCR using a spliced leader primer paired with an anchored oligo-dT primer (SL-1-nested and anchored oligo-dT; Supplemental Table 1). 20 cycles of PCR were completed (55°C annealing, 45s extension) using Phusion polymerase (Thermo Scientific, #F530L). PCR reactions were cleaned up with 1.8x Mag-Bind beads. After amplifying *T. brucei* cDNA, a VSG-specific PCR reaction was carried out using M13RSL and 14-mer-SP6 primers (see primers; Supplemental Table 1). 30 cycles of PCR (42°C annealing, 45s extension) were performed using Phusion polymerase. Amplified VSG cDNA was then cleaned up with 1X Mag-Bind beads and quantified using a Qubit dsDNA HS Assay (Invitrogen Q32854).

Sequencing libraries were prepared from 1 ng of each VSG PCR product using the Nextera XT DNA Library Preparation Kit (Illumina, FC-131-1096) following the manufacturer's protocol except for the final cleanup step, which was performed using 1X Mag-Bind beads. Single-end 100bp sequencing was performed on an Illumina HiSeq 2500. Raw data are available in the National Center for Biotechnology Information (NCBI) Sequence Read Archive under accession number PRJNA751607.

VSG-seq analysis of *T. b. gambiense* and *T. b. rhodesiense* sequencing libraries

To analyze both *T. b. gambiense* (VSG-seq preparations) and *T. b. rhodesiense* (traditional mRNA sequencing library preparations; sequences were obtained from ENA, accession numbers PRJEB27207 and PRJEB18523), we processed raw reads using the VSG-seq pipeline available at <https://github.com/mugnierlab/VSGSeqPipeline>.

Briefly, VSG transcripts were assembled *de novo* from quality- and adapter-trimmed reads for each sample (patient or patient replicate) from raw reads using Trinity (version 5.26.2) [43]. Contigs containing open reading frames (ORFs) were identified as previously described [30]. ORF-containing contigs were compared to Lister 427 and EATRO1125 VSGs as well as a collection of known contaminating non-VSG sequences. Alignments to VSGs with an E-value below 1×10^{-10} that did not match any known non-VSG contaminants were identified as VSG transcripts. For *T. b. gambiense* replicate libraries, VSG ORFs identified in any patient replicate were consolidated into a sole reference genome for each patient using CD-HIT (version 4.8.1)[44] with the following arguments: -d 0 -c 0.98 -n 8 -G 1 -g 1 -s 0.0 -aL 0.0. Final VSG ORF files were manually inspected.

Two *T. b. gambiense* patient VSGs (Patients 11 and 13) showed likely assembly errors. In one case, a VSG was duplicated and concatenated, and in another, two VSGs were concatenated. These reference files were manually corrected (removing the duplicate or editing annotation to reflect two VSGs in the concatenated ORF) so that each VSG could be properly quantified. VSG reference databases for each patient are available at <https://github.com/mugnierlab/Tbgambiense2021/>. For *T. b. gambiense*, we then aligned reads from each patient replicate to that patient's consolidated reference genome using Bowtie with the parameters -v 2 -m 1 -S (version 1.2.3) [45].

For *T. b. rhodesiense*, we aligned each patient's data to its own VSG ORF assembly. RPKM values for each VSG in each sample were generated using MULTo (version 1.0) [46], and the percentage of parasites in each population expressing a VSG was calculated as described previously [30]. For *T. b. gambiense* samples, we included only VSGs with an expression measurement above 1% in two or more patient replicates in our analysis. For *T. b. rhodesiense* samples, we included only VSGs with expression >0.01%. To compare VSG expression between patients, despite the different reference genomes used for each patient, we used CD-HIT to cluster VSG sequences with greater than 98% similarity among patients, using the same parameters used to consolidate reference VSG databases before alignment. We gave each unique VSG cluster a numerical ID (e.g., Gambiense #) and chose the longest sequence within each group to represent the cluster. Before analysis, we manually removed clusters representing TgsGP and SRA from the expressed VSG sets. UpSet plots were made with the UpSetR package[47]. The R code used to analyze resulting data and generate figures is available at <https://github.com/mugnierlab/Tbgambiense2021/>.

Analysis of VSG N-terminal domains

Genomic VSG sequences

The VSG repertoires of *T. b. brucei* Lister 427 (“Lister427_2018” assembly), *T. b. brucei* TREU927/4 and *T. b. gambiense* DAL972 were taken from TriTrypDB (v50), while the *T. b. brucei* EATRO 1125 VSGnome was used for analysis of the EATRO1125 VSG repertoire (vs_gsb1125_nodups_atleast250aas_pro.txt, available at <https://tryps.rockefeller.edu/Sequences.html> or GenBank accession KX698609.1 - KX701858.1). VSG sequences from other strains (except those generated by VSG-seq) were taken from the analysis in Cross, et al. [15]. Likely VSG N-termini were identified as predicted proteins with significant similarity (e-value $\leq 10^{-5}$) to hidden Markov models (HMMs) of aligned type A and B VSG N-termini taken from [15].

N-terminal domain phylogenies

Phylogenies of VSG N-termini based on unaligned sequence similarities were constructed using the method described in [48] and used previously to classify VSG sequence [15]. We extracted predicted N-terminal domain protein sequences by using the largest bounding envelope coordinates of a match to either type A or type B HMM. A matrix of similarities between all sequences was constructed from normalized transformed BLASTp scores as in Wickstead, et al. [48] and used to infer a neighbor-joining tree using QuickTree v1.1 [49]. Trees were annotated and visualized in R with the package APE v5.2 [50].

HMM

For N-terminal typing by HMM, we used a python analysis pipeline available at (https://github.com/mugnierlab/find_VSG_Ndomains). The pipeline first identifies the boundaries of the VSG N-terminal domain using the type A and type B HMM profiles generated by Cross *et al.* which includes 735 previously-typed VSG N-terminal domain sequences [15]. N-terminal domains are defined by the largest envelope domain coordinate that meets e-value threshold (1×10^{-5} , --domE 0.00001). In cases where no N-terminal domain is identified using these profiles, the pipeline executes a “rescue” domain search in which the VSG is searched against a ‘pan-VSG’ N-terminus profile we generated using 763 previously-typed VSG N-terminal domain sequences. This set of VSGs includes several *T. brucei* strains and/or subspecies: Tb427 (559), TREU927 (138), *T. b. gambiense* DAL972 (28), EATRO795 (8), EATRO110 (5), *T. equiperdum* (4), and *T. evansi* (21). The N-terminal domain type of these VSGs were previously determined by Cross et. al (2014) by building neighbor-joining trees using local alignment scores from all-versus-all BLASTp similarity searches [15]. Domain boundaries are called using the same parameters as with the type A and B profiles.

After identifying boundaries, the pipeline extracts the sequence of the N-terminal domain, and this is searched against five subtype HMM profiles. To generate N-terminal domain subtype HMM profiles, five multiple sequence alignments were performed using Clustal Omega [51] with the 763 previously-typed VSG N-terminal domain sequences described above; each alignment included the VSG N-terminal domains of the same subtype (A1, A2, A3, B1, and B2). Alignment output files in STOCKHOLM format were

used to generate distinct HMM profiles for type A1, A2, A3, B1, and B2 VSGs using the pre-determined subtype classifications of the 763 VSGs using HMMer version 3.1b2 [52]. The number of sequences used to create each subtype profile ranged from 75 to 211. The most probable subtype is determined by the pipeline based on the highest scoring sequence alignment against the subtype HMM profile database when HMMscan is run under default alignment parameters. The pipeline generates a FASTA file containing the amino acid sequence of each VSG N-terminus and a CSV with descriptions of the N-terminal domain including its type and subtype.

Network graph

N-terminal network graphs were made using VSG N-terminal domains from the TriTrypDB Lister427_2018 and *T. b. gambiense* DAL972 (v50) VSG sets described above, and the *T. b. gambiense* and *T. b. rhodesiense* patient VSG N-termini which met our expression thresholds. Identified N-terminal domains were then subjected to an all-versus-all BLASTp. A pairwise table was created that includes each query-subject pair, the corresponding alignment E-value, and N-terminal domain type of the query sequence if previously typed in Cross, et al. [15]. Pseudogenes and fragments were excluded from the Lister427_2018 reference prior to plotting by filtering for VSG genes annotated as pseudogenes and any less than 400 amino acids in length, as the remaining sequences are most likely to be full length VSG. Network graphs were generated with the igraph R package[53] using undirected and unweighted clustering of nodes after applying link cutoffs based on E-value $< 10^{-2}$. The leading eigenvector clustering method [54] was used to detect and assign nodes to communities based on clustering (cluster_leading_eigen() method in igraph).

Analysis of VSG C-terminal domains

VSG C-termini were extracted from expressed *T. b. gambiense* VSGs, *T.b. gambiense* DAL972 (v50), and 545 previously-typed VSG C-termini from the Lister 427 strain using the C-terminal HMM profile generated by Cross et al.[15] and the same HMMscan parameters as for N-termini (E-value $< 1 \times 10^{-5}$; largest domain based on envelope coordinates). An all-vs-all BLASTp was performed on these sequences, and network graphs were generated in the same manner as the N-terminal network graphs. Links were drawn between C-termini with a BLASTp E-value 1×10^{-3} . The leading eigenvector method for clustering [54] was used to detect and assign nodes to communities based on clustering (cluster_leading_eigen() method in igraph).

Comparison of gHAT patient VSGs to the DAL972 reference

The nucleotide sequences of expressed *T. b. gambiense* patient VSGs were compared to the complete *T. b. gambiense* DAL972 reference genome sequence (v54) from TriTrypDB using BLASTn. To find evidence of VSG diversification by mosaicism or segmental gene conversion, permissive parameters were used so short regions of homology would be reported (blastn -word_size 11 -evalue 0.1). Both query and subject alignment coordinates were recorded in the output for downstream analysis, along with

384 E-value and percent identity. Alignment ranges were plotted with the IRanges R
385 package[55].

Results

Parasites in gHAT patients express diverse sets of VSGs

To investigate VSG expression in natural human infections, we performed VSG-seq on RNA extracted from whole blood collected from 12 human African trypanosomiasis patients from five locations in the Kwilu province of DR Congo (Figure 1A). We estimated the relative parasitemia of each patient by SL-QPCR [56], and we estimated the number of parasites after mAECT on buffy coat for all patients except patient 29 (Table 1). Using RNA extracted from 2.5 mL of whole blood from each patient, we prepared libraries for VSG-seq in three separate batches for each technical replicate. We amplified *T. brucei* RNA from host/parasite total RNA using a primer against the *T. brucei* spliced leader sequence and an anchored oligo-dT primer. The resulting trypanosome-enriched cDNA was used as a template to amplify VSG cDNA, and VSG amplicons were then submitted to VSG-seq sequencing and analysis. To determine whether a VSG was expressed within a patient, we applied the following stringent cutoffs:

- 1) We conservatively estimate that each 2.5 mL patient blood sample contained a minimum of 100 parasites. At this minimum parasitemia, a single parasite would represent 1% of the population (and consequently ~1% of the parasite RNA in a sample). As a result, we excluded all VSGs expressed by <1% of the population as estimated by VSG-seq.
- 2) We classified VSGs as expressed if they met the expression cutoff in at least two of three technical library replicates.

1112 unique VSG open reading frames were assembled *de novo* in the patient set and 44 met our expression criteria. Only these 44 VSGs, which we will refer to as “expressed VSGs,” were considered in downstream analysis, except when otherwise noted. TgsGP, the VSG-like protein which partially enables resistance to human serum in *T. b. gambiense* [57], assembled in samples from patients 2, 11, 13, and 17, and met the expression threshold in patients 2, 11, and 17. The absence of this transcript in most samples is likely due to the low amount of input material used to prepare samples. Notably, none of the expressed VSGs shared significant nucleotide sequence similarity outside the conserved C-terminal domain with any VSGs in the *T. b. gambiense* DAL972 genome (Supplemental Figures 6 and 7).

At least one VSG met our expression criteria in each patient, and in most cases, multiple VSGs were detected. Patient 2 showed the highest diversity, with 14 VSGs expressed (Figure 1B, Supplemental Figure 1). We observed a correlation between parasitemia, as estimated by qPCR, and the number of VSGs (Supplemental Figure 2), suggesting that these samples do not reflect the full diversity of each population. Nevertheless, two VSGs were shared between patients: VSG ‘Gambiense 195’ was expressed in both patient 12 and patient 17 from Village C; VSG ‘Gambiense 38’ was expressed in patient 12 from Village C and patient 23 from Village D (Figure 1C). Because our sampling did not reach saturation, resulting in some variability between

technical replicates, we chose to focus only on the presence/absence of individual VSGs rather than relative expression levels within each population.

Patient	Location	est. parasites in 500µL buffy coat	mean SL-RNA Ct	WBC	Parasites in CSF	Stage
1	Village A	>50	22.155	1	-	First
2	Village A	>50	19.020	6	-	Early 2nd
3	Village A	2-5	28.780	6	-	Early 2nd
11	Village C	>50	22.030	9	-	Early 2nd
12	Village C	6-20	25.430	6	-	Early 2nd
13	Village C	6-20	26.635	12	-	Early 2nd
17	Village C	21-50	24.495	13	-	Early 2nd
19	Village C	1	28.245	7	-	Early 2nd
23	Village D	6-20	27.085	2	-	First
29	Village B	-	28.320	3	-	First
30	Village B	>50	22.960	694	+	Severe 2nd
33	Village E	1	32.385	2	-	First

Table 1. Patient stage and parasitemia data. We used the following staging definitions: First: 0-5 WBC/µl, no trypanosomes in cerebrospinal fluid (CSF). Second: >5 WBC/µl or trypanosomes in CSF (with early 2nd: 6-20 WBC/µl and no trypanosomes in CSF; severe 2nd: >100 WBC/µl). WBC: white blood cells.

Figure 1. Parasites isolated from gHAT patients express multiple VSGs. (A) Map showing the location of each patient's home village. Maps were generated with ArcGIS® software by Esri, using world imagery and National Geographic style basemaps. (B) Graph depicting the total number of VSGs expressed in each patient. (C) The intersection of expressed VSG sets in each patient. Bars on the left represent the size of the total set of VSGs expressed in each patient. Dots represent an intersection of sets with bars above the dots representing the size of the intersection. Color indicates patient origin.

Natural *T. b. gambiense* infections show a strong bias towards the expression of type B VSG

To further characterize the set of expressed VSGs in these samples, we sought a fast and unbiased method for determining the type and subtype of each VSG's N-terminal domain. We evaluated two approaches. The first approach was to create a bioinformatic pipeline to determine each N-terminal domain subtype, using HMM profiles we created for each subtype from sets of N-terminal domains previously typed by Cross et al. [15]. The second approach was to create a BLASTp network graph based on a published method [58] where the N-terminal subtype of a VSG is determined by the set of VSGs it clusters with, and clusters are identified using the leading eigenvector method [54]. We used each approach to determine the N-terminal subtype of each expressed VSG in our patient sample dataset, along with 863 VSG N-termini from the Lister 427 genome. We compared these results to either existing N-terminal classification (for Lister 427 VSGs) or classification based on position in a newly-generated BLASTp-tree [15] (for *T. b. gambiense* VSGs; Figure 2A).

Both the new HMM profile and BLASTp network graph approaches generally recapitulated previous VSG classification based on BLASTp-tree, with all three methods agreeing 93.7% of the time (Figure 2B). The HMM pipeline method agreed with BLASTp-tree typing for all patient VSGs, while the network graph approach agreed for 43/44 VSGs (Figure 2B, Supplemental Figure 3, Supplemental Table 1) [15]. It is not surprising that the HMM pipeline would better reflect the results of the BLASTp-tree method, as the N-terminal subtype HMM profiles were generated using VSGs classified by this method. Based on these data, we determined that the HMM method is a fast and accurate approach for determining the N-terminal domain types of unknown VSGs.

Our N-terminal domain typing pipeline identified the domain sequence and subtype for all 44 patient VSGs (Figure 2C). Of the expressed *T. b. gambiense* VSGs, 82% had type B N-terminal domains, and 50% or more of expressed VSGs within each patient were type B. This bias was not restricted to highly expressed VSGs, as 74.5% of all assembled VSG (813 of 1091 classifiable to an N-terminal subtype) were also type B. This observation motivated further investigation into the expressed N-terminal domains in infections by other *T. brucei* subspecies.

Figure 2. *T. b. gambiense* samples show a bias towards the expression of type B VSG. (A) Visualization of relatedness between N-terminal domain peptide sequences inferred by Neighbor-Joining based on normalized BLASTp scores. Legend indicates classification by HMM pipeline (for Lister 427 VSGs, to highlight agreement between the two methods) or by subspecies for VSGs expressed in patients. (B) Agreement between three VSG typing methods for Lister 427 VSG set and the expressed *T. b. gambiense* patient VSG set. (C) N-terminal domain subtype composition of expressed *T. b. gambiense* VSGs as determined by HMM analysis pipeline.

Type B VSG bias is not observed in natural *T. b. rhodesiense* infection

To determine whether the bias towards type B VSGs was unique to *T. b. gambiense* infections, we analyzed RNA-seq data from a published study measuring gene expression in the blood and cerebrospinal fluid (CSF) of *T. b. rhodesiense* patients in Northern Uganda [59]. These libraries were prepared conventionally after either rRNA-depletion for blood or poly-A selection for CSF samples. We analyzed only those samples for which at least 10% of reads mapped to the *T. brucei* genome. Raw reads from these samples were subjected to the VSG-seq analysis pipeline. Because the estimated parasitemia of these patients was much higher than our *T. b. gambiense* study, we adjusted our expression criteria accordingly to 0.01%, the published limit of detection of VSG-seq [30]. Using this approach, we identified 77 unique VSG sequences across all blood and CSF samples (Figure 3A, Supplemental Figure 4). SRA, the VSG-like protein that confers human serum resistance in *T. b. rhodesiense* [60], was expressed in all patient samples.

The HMM pipeline determined types for 74 of these VSG sequences; the remaining sequences appeared to be incompletely assembled, presumably due to insufficient read depth from their low level of expression. Multiple VSGs assembled in each patient (Figure 3A), and a large proportion of VSGs were expressed in multiple patients (Figure 3C), in line with our observations in experimental mouse infections. Although most VSGs detected in these patients were type B (57%, Figure 3B), this VSG type was much less predominant than in *T. b. gambiense* infection. Interestingly, *T. b. rhodesiense* patient CSF revealed another possible layer of diversity in VSG expression, with 5 VSGs expressed exclusively in this space.

Figure 3. *T. b. rhodesiense* samples reveal diverse VSG expression but little N-terminal type bias. (A) The total number of expressed *T. b. rhodesiense* VSGs in each patient and sample type. Bar color represents the sample type from which RNA was extracted. (B) N-terminal domain subtype composition of all expressed VSGs. (C) Intersections of VSGs expressed in multiple infections. Bars on the left represent the size of the total set of VSGs expressed in each patient. Dots represent an intersection of sets, with bars above the dots representing the size of the intersection. Color indicates patient origin.

The predominant VSG N-terminal type fluctuates over time during experimental *T. b. brucei* infection

One explanation for the bias towards type B VSG in *T. b. gambiense* could be that VSG type stabilizes over time and type B predominates at later infection timepoints. This explanation is plausible because patient samples only represent a single moment during infection, and *T. b. gambiense* samples are more likely to be obtained at much later timepoints post-infection. To investigate whether N-terminal type predominance is stable or fluctuates over the course of infection, we took advantage of our published VSG-seq analysis of parasites isolated from mice infected with the *T.b. brucei* EATRO1125 strain. Blood was collected over time during this study, providing data from days 6/7, 12, 14, 21, 24, 26, and 30 post-infection in all four mice, and from days 96, 99, 102, and 105 in one of the four mice (Mouse 3). Of 192 unique VSGs identified between days 0-30, the python HMM pipeline typed 190; of 97 unique VSGs identified between days 96-105, the pipeline typed 93 VSGs. The remaining VSGs were incompletely assembled by Trinity. Our analysis of VSG types over time revealed that the predominantly expressed N-terminal domain type fluctuates between type A and type B throughout the early stages of infection and in extended chronic infections (Figure 4). Parasitemia did not correlate with either the diversity of VSG expression or N-terminal domain type predominance (Supplemental Figure 3). Because all patient samples were collected from a single timepoint in varying stages of disease, it remains unclear whether the predominant N-terminal type continues to fluctuate in human infections over time as it does in mice that have reached chronic stages of infection.

Figure 4. VSG N-terminal type composition fluctuates over the course of infection in mice. Proportions of N-terminal domain types expressed in *T. b. brucei* infected mice over time. The black dotted line represents the total number of identified VSGs. A) N-terminal type composition days 0-30. B) Type composition days 96-105.

The composition of the genomic VSG repertoire is reflected in expressed VSG N-terminal domain types

Another source for bias in expressed VSG type is the composition of the genomic VSG repertoire. We could directly compare the genomic and expressed VSG repertoire for EATRO1125 mouse infections, as the 'VSGnome' for this strain has been sequenced. Although the predominant N-terminal VSG type fluctuates throughout these infections, the expressed VSG repertoire across all time points generally reflects the composition of the genomic repertoire (chi-squared $p = 0.0515$, Figure 5A).

Unfortunately, the entire repertoire of VSGs encoded by most trypanosome strains is unknown, so such a direct comparison is impossible for *T. b. gambiense* and *T. b. rhodesiense* patient samples. Although the content of the 'core' *T. brucei* genome (containing the diploid, housekeeping genes) is similar enough among subspecies for resequencing projects to be scaffolded using the TREU927 or Lister 427 reference genomes [61–63], it is not clear whether the VSG repertoires of subspecies (or even individual parasite strains [36]) share this degree of similarity. Because no near-complete VSGnome for any *T. b. rhodesiense* strain was available, we compared the makeup of *T. b. rhodesiense* expressed VSGs with the closely related and near-complete *T. b. brucei* Lister 427 repertoire [62]. We observed no difference in the proportions of N-terminal types (chi-squared p -value = 0.2422) (Figure 5B). Similarly, the proportion of N-terminal domains identified in the *T. b. gambiense* patient samples is not statistically different from the incomplete *T. b. gambiense* DAL972 genomic repertoire (chi-squared p -value = 0.0575) (Figure 5B). Both *T. b. gambiense* patient VSGs (chi-squared p -value = 2.413×10^{-4}) and the 54 VSGs identified in *T. b. gambiense* DAL972 (chi-square p -value = 0.0301) have A and B type frequencies that differ significantly from Lister427. These results should be interpreted with caution given the limitations of the reference genomes used. Taken together, however, they support a model in which VSG types are drawn from the repertoire at a roughly equal frequency to their representation in the genome, with *T. b. gambiense* exhibiting an N-terminal type composition that differs from other subspecies.

Figure 5. VSG expression reflects the genomic VSG repertoire of the infecting parasites.

(A) Columns show the proportion of VSG types identified in each mouse infection over all time points and the proportion of VSG types in the infecting *T. b. brucei* strain, EATRO 1125. The total number of unique VSG sequences is displayed above each column. (B) A comparison of the frequencies of type A and B VSGs expressed in patients and those present in Lister 427 and DAL972 reference genomes. Relevant statistical comparisons are shown.

***T. b. gambiense* expressed VSG C-terminal domains show a bias for type 2 domains**

In addition to examining N-terminal types in our *T. b. gambiense* dataset, we also examined expressed VSG C-termini. Previous studies defined six C-terminal types, although resolving these types in larger VSG sets has been difficult due to the extensive similarity among VSG C-termini [15,58]. In line with previous observations, a BLASTp-tree analysis of assembled *T. b. gambiense* C-terminal domains revealed frequent sequence similarity between expressed C-terminal types but did not provide sufficient resolution to confidently assign types (Fig 6A). To supplement this analysis, we also performed a network graph analysis. Although this method had previously performed poorly in resolving VSG C-termini [58], using the leading eigenvector clustering method [54] to define community membership within the graph allowed a faithful reconstitution of the C-terminal types previously determined by BLASTp-tree analysis.

Using the network graph approach, we tentatively assigned C-terminal domain types to the *T. b. gambiense* VSGs (Figure 6B). Most patient C-terminal domain types were type 2, while the remaining types were predominantly type 1. Only one type 3 C-terminus was identified in the patient set. Although there are very few VSG C-termini available in the *T. b. gambiense* DAL972 genome, these sequences show a predominance of types 3 and 5. In contrast, the genomic repertoire of C-termini in Lister 427 shows a roughly equivalent representation of types 1, 2, and 3. Unlike *T. b. gambiense* N-termini, the expressed *T. b. gambiense* C-termini were more restricted than the sets of C-termini apparently available in the *T. b. gambiense* DAL972 or *T. b. brucei* Lister 427 genomes (Fisher's exact test, $p\text{-value} < 1 \times 10^{-5}$).

In line with previous observations, we saw no evidence of domain exclusion: a C-terminal domain of one type could be paired with any type of N-terminal domain (Figure 6E)[20]. As observed in our analysis of expressed N-termini, C-terminal domain types were not correlated to geographical origin (Figure 6F). Overall, these data suggest that, like N-termini, expressed VSG C-termini are also biased towards certain C-terminal types. Unlike N-termini, however, C-terminal types expressed in *T. b. gambiense* infection may not reflect the parasite's genomic repertoire composition.

Figure 6. Expressed VSG C-termini are primarily type 1 and type 2. A) BLASTp-tree of C-terminal domains. Points are colored based on previously determined C-terminal type from Cross et al. or by the source of the sequence (genomic or expressed) for *T. b. gambiense* VSGs. B) Network plot showing peptide sequence relatedness between C-terminal domains in *T. b. gambiense* expressed VSGs. Each point represents a VSG C-terminus. A link was drawn between points if the BLASTp e-value was less than 1×10^{-3} . Points are colored by the cluster determined by the clustering algorithm. Shaded circles also indicate clusters. C) Same network plot as in B but colored by previously determined C-terminal type from Cross et al., or by source for unclassified genomic or expressed VSGs. D) VSG C-terminal types, based on cluster assignment visualized in panel B, in genomic and expressed VSG sets. E) Pairing of C- and N-termini in *T. b. gambiense* patients. F) C-termini detected in each patient village.

Discussion

African trypanosomes evade the host adaptive immune response through a process of antigenic variation where parasites switch their expressed VSG [64]. The genome of *T. brucei* encodes a large repertoire of VSG genes, pseudogenes, and gene fragments that can be expanded continuously through recombination to form entirely novel “mosaic” VSGs [17]. While antigenic variation has been studied extensively in culture and animal infection models, our understanding of the process in natural infections, particularly human infection, is limited. Most experimental mouse infections are sustained for weeks to months, while humans and large mammals may be infected for several months or even years. Additionally, laboratory studies of antigenic variation almost exclusively use *T. b. brucei*, a subspecies of *T. brucei* that, by definition, does not infect humans.

The primary hurdle to exploring antigenic variation in nature has been technical: it is difficult to obtain sufficient parasite material for analysis. This is especially true for infection with *T. b. gambiense*, which often exhibits extremely low parasitemia. VSG-seq, which relies on PCR and requires very small amounts of material for analysis, provides a new tool for exploring VSG expression in natural human infections. Here we have demonstrated the feasibility of VSG-seq to analyze VSG expression in RNA samples isolated directly from HAT patients. Our analyses show that the diversity seen in mouse infection models is mostly recapitulated in natural infection; however, there may be unique aspects of antigenic variation in *T. b. gambiense* that can only be explored by studying natural infections.

The most notable result in our previous analysis of mouse infections was the diversity of VSGs expressed within each infection. Rather than a few VSGs expressed at a time, we saw many VSGs expressed simultaneously in each parasite population. This finding confirms previous estimates of antigenic diversity in experimental mouse infection [17] and suggests that the genomic VSG repertoire might be exploited very rapidly. The study presented here detected several expressed VSGs in most *T. b. gambiense* samples. Although diversity in *T. b. gambiense* infection appeared lower overall, the correlation we observed between parasitemia and diversity in *T. b. gambiense* isolates suggests that our sampling was incomplete. Indeed, in our analysis of *T. b. rhodesiense* infection (a more reasonable comparison to mouse infection given similar expression cutoffs and parasitemia), we observed diversity similar to or higher than what we have detected in *T. b. brucei* mouse infections. Moreover, *T. b. rhodesiense* patient CSF revealed another layer of diversity in VSG expression, with 5 VSGs expressed exclusively in this space. Although this observation is difficult to interpret without information about the precise timing of sample collection, it is exciting to speculate that different organs or body compartments could harbor different sets of VSGs. Overall, our analysis of VSG expression in *T. b. gambiense* and *T. b. rhodesiense* patients confirmed the long-held assumption that VSG diversity is a feature of natural infection.

While analyzing the sets of expressed VSGs in *T. b. gambiense* and *T. b. rhodesiense* infections, we found evidence for another feature of experimental infection that holds

true in a natural host: hierarchical VSG expression. Both *in vitro* and *in vivo* studies have shown that VSG switching is not entirely stochastic but rather hierarchical. Specific variants dominate expression in the parasite population in a reproducible order that appears to be independent of the starting VSG [17,31,32,65,66]. In the *T. b. gambiense* samples, we found two VSGs that met our detection threshold in multiple patients. In *T. b. rhodesiense* samples, multiple patients shared a large proportion of expressed VSGs. Given the large size of the genomic VSG repertoire, any overlap in expressed VSG repertoire is likely indicative of a semi-predictable hierarchy of switching and preference for the expression of certain VSGs.

Of the two shared VSGs we identified in the *T. b. gambiense* patients, one was identified in two patients from the same village, while the other was found in two patients from villages 40km apart. At this short distance, it is possible that the infecting parasites were genetically similar. However, it would be interesting to investigate whether preference for the expression of certain VSGs occurs even between parasites isolated at greater distances. Indeed, past research has shown that the sensitivity of serological tests for gHAT, which detect antibodies against the LiTat 1.3 VSG, vary regionally, potentially due to differences in the underlying genomic or expressed VSG repertoire in circulating strains [67,68].

Our data are consistent with the notion that there could be geographic variations in *T. brucei* VSG repertoires. None of our assembled *T. b. gambiense* VSGs from patients in the DR Congo shared significant N-terminal sequence similarity with those in the genome of DAL972, a parasite isolate from Côte d'Ivoire. In contrast, the C-termini of patient VSGs were well represented by similar sequences within the DAL972 reference. Geographic variation has been observed in *var* gene repertoires of *Plasmodium falciparum* [69] and the VSG repertoire of *Trypanosoma vivax*, another African trypanosome [35]. A better understanding of such differences in *T. brucei*, if they exist, could lead to development of more sensitive HAT diagnostics.

To better understand the VSG proteins expressed in natural infections, we developed an HMM VSG typing pipeline that revealed an intriguing bias in *T. b. gambiense* infection towards type B VSGs, which appears to be specific to *T. b. gambiense* patient samples. Small sample sizes and important differences between each *T. brucei* subspecies' dataset limit the conclusions that can be drawn. However, comparisons between these sets suggest that the genomic VSG repertoire determines the distribution of VSG N-terminal types expressed during *T. brucei* infection and may account for the bias we have observed in *T. b. gambiense* patients. Thus, the *T. b. gambiense* VSG repertoire may contain a larger proportion of type B VSGs than its more virulent counterparts.

Another possibility we cannot rule out, however, is that the gHAT samples are biased due to selection by the serological test used for diagnosis. Patients were screened for *T. b. gambiense* infection using the CATT, a serological test that uses parasites expressing VSG LiTat 1.3 as an antigen. LiTat 1.3 contains a type B2 N-terminal domain [67,68]. Patients infected with parasites predominantly expressing type B VSGs

may be more likely to generate antibodies that cross-react with LiTat1.3, resulting in preferential detection of these cases. In contrast, *T.b. rhodesiense* can only be diagnosed microscopically, removing the potential to introduce bias through screening. It remains to be investigated whether samples from patients diagnosed using newer screening tests, which include the invariant surface glycoprotein ISG65 and the type A VSG LiTat 1.5 [29], would show similar bias towards the expression of type B VSGs.

Analysis of expressed VSG C-terminal domains in *T. b. gambiense* patients showed a bias towards C-terminal types 1 and 2. The diagnostic VSG LiTat1.3 contains a type 3 C-terminus, a C-terminal type underrepresented in the patient set. Therefore, it is unlikely that a bias in expressed C-terminal types is related to the screening method. Notably, the bias toward C-terminal types 1 and 2 was not reflected in the limited VSG repertoire of the DAL972 reference genome or the repertoire of the Lister 427 *T. b. brucei* reference genome. This finding could be related to the limited set of VSGs present in the DAL972 reference genome or reflect a true expression bias.

Could a bias towards certain VSG types, whether due to a difference in repertoire composition or expression preference, account for unique features of *T. b. gambiense* infection, including its chronicity and primarily anthroponotic nature [70]? While the genomic VSG repertoire has been analyzed extensively in laboratory strains, little is known about how differences in VSG proteins relate to parasite biology or whether there could be biological consequences to the expression of specific VSG N- or C-terminal types. Type A *var* genes in *Plasmodium falciparum* infection are associated with severe malaria [71–75], and similar mechanisms have been hypothesized to exist in *T. vivax* and *T. congolense* infections [33,35,76,77].

In *T. brucei*, several VSGs have evolved specific functions besides antigenic variation [77]. The first type B VSG structure was recently solved [78], revealing a unique O-linked carbohydrate in the VSG's N-terminal domain. This modification interfered with the generation of protective immunity in a mouse infection model; perhaps structural differences between each VSG type, including glycosylation patterns, could influence infection outcomes. Further research will be needed to determine whether the observed predominance of type B VSGs could influence the biology of *T. b. gambiense* infection.

Currently, it is unclear why this collection of gHAT patient isolates demonstrates a bias towards the expression of certain VSG types. More research will be needed to determine whether the *T. b. gambiense* VSG repertoire contains a unique distribution of VSG types, whether these parasites preferentially express certain VSG types, and whether this bias could have functional consequences. This study has shown, however, is that it is feasible to explore antigenic variation in natural infection. Moreover, although mouse models do reflect the general dynamics of antigenic variation in natural *T. brucei* infection, unique biology remains to be uncovered by studying this process in its natural context.

Acknowledgments

We are very grateful to the patients without whom this work would not have been possible. We thank George Cross and Danae Schulz for comments on the manuscript, and Mary Gebhardt for help with GIS. The Atlas of HAT is an initiative of the World Health Organization (WHO), jointly implemented with the Food and Agriculture Organization of the United Nations (FAO) in the framework of the Programme Against African Trypanosomiasis (PAAT). Field work and specimen collection in DR Congo were funded through the Wellcome Trust (study number 099310/Z/12/Z) awarded to the TrypanoGEN Consortium (www.trypanogen.net), members of H3Africa (h3africa.org). Sample work-up was supported by the Research Foundation Flanders (FWO grant 1501413N). Work by BW was supported by University of Nottingham/Wellcome Trust Institutional Strategic Support Fund award 204843/Z/16/Z. MRM and SS were supported by Office of the Director, NIH (DP5OD023065). JS is supported by NIH T32AI007417.

Supplement

Supplemental Table 1. Primer sequences.

Supplemental Table 2. gHAT patient distance matrix.

Supplemental Table 3. gHAT VSG expression data.

Supplemental Table 4. Tables comparing BLAST-tree, HMMscan, and network plot typing methods.

Supplemental Table 5. rHAT VSG expression data.

Supplemental Figure 2. Correlation between parasitemia and diversity and N-terminal type distribution. (A) Correlation plots for VSG diversity and percent of N-terminal domain type B for *T.b. brucei* infected mice from Mugnier et al. 2015. (B) Correlation plots for *T.b. rhodesiense* infected patients from Mulindwa et al. 2018. (C) Correlation plots for *T.b. gambiense* infected patients.

Supplemental Figure 3. (A) Network plot showing peptide sequence relatedness between N-terminal domains. Each point represents a VSG N-terminus. A link was drawn between points if the BLASTp e-value was less than 10^{-2} . Colors and shaded circles represent community assignments determined by the clustering algorithm. (B) The same graph as in (A), but points are manually colored by known N-terminal subtype from Cross et al. or by subspecies for VSGs identified in patients.

846

847 **Supplemental Figure 4. BLASTp-tree of all *T. b. gambiense* VSGs.** File attached.

848

849

Supplemental Figure 5. Heatmap of all assembled *T. b. rhodesiense* patient VSGs. Greyscale shows log_{10} of the estimated percentage of the parasite population expressing each VSG. Variants expressed by less than 0.01% of parasites considered not detected (n.d.).

Supplemental Figure 6. Ranges of all BLASTn hits with e-value < 0.1 for each assembled gHAT patient VSG against the TbgDAL972 whole genome reference. The full length of each patient VSG in base pairs is plotted as the X-axis of each plot facet. Bars color-coded by percent identity show each hit's alignment length and position within the query VSG. Alignments within the same query coordinates that hit multiple locations in the reference genome are stacked in bins. The maximum number of bins is plotted on the y-axis.

Supplemental Figure 7. Patient VSG representation within TbgDAL972. A histogram of the percentage of each patient VSG that was represented anywhere within the Tbg DAL972 genome by BLASTn with a threshold E-value < 0.1.

References

1. Romero-Meza G, Mugnier MR. Trypanosoma brucei. Trends in Parasitology. 2019. doi:10.1016/j.pt.2019.10.007
2. Büscher P, Cecchi G, Jamonneau V, Priotto G. Human African trypanosomiasis. www.thelancet.com. 2017;390. doi:10.1016/S0140-6736(17)31510-6
3. Franco JR, Cecchi G, Priotto G, Paone M, Diarra A, Grout L, et al. Monitoring the elimination of human African trypanosomiasis at continental and country level: Update to 2018. PLoS Negl Trop Dis. 2020;14: e0008261. doi:10.1371/JOURNAL.PNTD.0008261
4. Kennedy PGE, Rodgers J. Clinical and neuropathogenetic aspects of human African trypanosomiasis. Frontiers in Immunology. Frontiers Media S.A.; 2019. doi:10.3389/fimmu.2019.00039
5. Franco JR, Cecchi G, Priotto G, Paone M, Diarra A, Grout L, et al. Monitoring the elimination of human African trypanosomiasis at continental and country level: Update to 2018. PLoS Negl Trop Dis. 2020;14: e0008261. doi:10.1371/JOURNAL.PNTD.0008261
6. WHO. The Global Health Observatory. In: WHO [Internet]. World Health Organization; [cited 17 Aug 2021]. Available: <https://www.who.int/data/gho/data/themes/topics/indicator-groups/indicator-group-details/GHO/human-african-trypanosomiasis>
7. WHO. Ending the neglect to attain the sustainable development goals: a road map for neglected tropical diseases 2021–2030. Geneva; 2020.
8. Magez S, Caljon G, Tran T, Stijlemans B, Radwanska M. Current status of vaccination against African trypanosomiasis. Parasitology. 2010;137: 2017–2027.
9. Trindade S, Rijo-Ferreira F, Carvalho T, Pinto-Neves D, Guegan F, Aresta-Branco F, et al. Trypanosoma brucei Parasites Occupy and Functionally Adapt to the Adipose Tissue in Mice. Cell Host Microbe. 2016;19: 837–848. doi:10.1016/j.chom.2016.05.002
10. Pereira SS, Trindade S, De Niz M, Figueiredo LM. Tissue tropism in parasitic diseases. Open Biology. Royal Society Publishing; 2019. doi:10.1098/rsob.190036
11. Camara M, Soumah AM, Ilboudo H, Travaillé C, Clucas C, Cooper A, et al. Extravascular Dermal Trypanosomes in Suspected and Confirmed Cases of gambiense Human African Trypanosomiasis. Clin Infect Dis. 2021;73: 12–20. doi:10.1093/cid/ciaa897
12. Alfituri OA, Bradford BM, Paxton E, Morrison LJ, Mabbott NA. Influence of the Draining Lymph Nodes and Organized Lymphoid Tissue Microarchitecture on Susceptibility to Intradermal Trypanosoma brucei Infection. Front Immunol. 2020;11. doi:10.3389/fimmu.2020.01118
13. Marcello L, Barry JD. Analysis of the VSG gene silent archive in Trypanosoma brucei reveals that mosaic gene expression is prominent in antigenic variation and is favored by archive substructure. Genome Res. 2007;17: 1344–1352.
14. Berriman M, Ghedin E, Hertz-Fowler C, Blandin G, Renauld H, Bartholomeu DC, et al. The genome of the African trypanosome Trypanosoma brucei. Sci (New York, NY). 2005;309: 416–422.
15. Cross GAM, Kim H-S, Wickstead B. Capturing the variant surface glycoprotein

- repertoire (the VSGnome) of *Trypanosoma brucei* Lister 427. *Mol Biochem Parasitol.* 2014;195: 59–73. doi:10.1016/j.molbiopara.2014.06.004
16. Müller LSM, Cosentino RO, Förstner KU, Guizetti J, Wedel C, Kaplan N, et al. Genome organization and DNA accessibility control antigenic variation in trypanosomes. *Nature.* 2018;563: 121–125.
17. Hall JPJ, Wang H, Barry JD. Mosaic VSGs and the scale of *Trypanosoma brucei* antigenic variation. Horn D, editor. *PLoS Pathog.* 2013;9: e1003502.
18. Jayaraman S, Harris C, Paxton E, Donachie A-M, Vaikkinen H, McCulloch R, et al. Application of long read sequencing to determine expressed antigen diversity in *Trypanosoma brucei* infections. Acosta-Serrano A, editor. *PLoS Negl Trop Dis.* 2019;13: e0007262. doi:10.1371/journal.pntd.0007262
19. Carrington M, Miller N, Blum M, Roditi I, Wiley D, Turner M. Variant specific glycoprotein of *Trypanosoma brucei* consists of two domains each having an independently conserved pattern of cysteine residues. *J Mol Biol.* 1991;221: 823–835.
20. Hutchinson OC, Smith W, Jones NG, Chattopadhyay A, Welburn SC, Carrington M. VSG structure: similar N-terminal domains can form functional VSGs with different types of C-terminal domain. *Mol Biochem Parasitol.* 2003;130: 127–131.
21. Jones NG, Nietlispach D, Sharma R, Burke DF, Eyres I, Mues M, et al. Structure of a glycosylphosphatidylinositol-anchored domain from a trypanosome variant surface glycoprotein. *J Biol Chem.* 2008;283: 3584–3593.
22. Schwede A, Jones N, Engstler M, Carrington M. The VSG C-terminal domain is inaccessible to antibodies on live trypanosomes. *Mol Biochem Parasitol.* 2011;175: 201–204.
23. Van Meirvenne N, Magnus E, Büscher P. Evaluation of variant specific trypanolysis tests for serodiagnosis of human infections with *Trypanosoma brucei gambiense*. *Acta Trop.* 1995;60: 189–199. doi:10.1016/0001-706X(95)00127-Z
24. Magnus E, Vervoort T, Van Meirvenne N. A card-agglutination test with stained trypanosomes (C.A.T.T.) for the serological diagnosis of *T. b. gambiense* trypanosomiasis. *Ann Soc Belg Med Trop (1920).* 1978;58: 169–176.
25. Dukes P, Gibson WC, Gashumba JK, Hudson KM, Bromidge TJ, Kaukus A, et al. Absence of the LiTat 1.3 (CATT antigen) gene in *Trypanosoma brucei gambiense* stocks from Cameroon. *Acta Trop.* 1992.
26. Enyaru JCK, Allingham R, Bromidge T, Kanmogne GD, Carasco JF. The isolation and genetic heterogeneity of *Trypanosoma brucei gambiense* from north-west Uganda. *Acta Trop.* 1993.
27. Bisser S, Lumbala C, Nguertoum E, Kande V, Flevaud L, Vatunga G, et al. Sensitivity and Specificity of a Prototype Rapid Diagnostic Test for the Detection of *Trypanosoma brucei gambiense* Infection: A Multi-centric Prospective Study. Acosta-Serrano A, editor. *PLoS Negl Trop Dis.* 2016;10: e0004608. doi:10.1371/journal.pntd.0004608
28. Büscher P, Mertens P, Leclipteux T, Gillemann Q, Jacquet D, Mumba-Ngoyi D, et al. Sensitivity and specificity of HAT Sero-K-SeT, a rapid diagnostic test for serodiagnosis of sleeping sickness caused by *Trypanosoma brucei gambiense*: A case-control study. *Lancet Glob Heal.* 2014;2: e359–e363. doi:10.1016/S2214-109X(14)70203-7

- 960 29. Lumbala C, Biéler S, Kayembe S, Makabuza J, Ongarello S, Ndung'u JM.
 961 Prospective evaluation of a rapid diagnostic test for *Trypanosoma brucei*
 962 gambiense infection developed using recombinant antigens. *PLoS Negl Trop Dis*.
 963 2018;12: e0006386. doi:10.1371/JOURNAL.PNTD.0006386
- 964 30. Mugnier MR, Cross GAM, Papavasiliou FN. The in vivo dynamics of antigenic
 965 variation in *Trypanosoma brucei*. *Science* (80-). 2015;347: 1470–1473.
 966 doi:10.1126/science.aaa4502
- 967 31. Morrison LJ, Majiwa P, Read AF, Barry JD. Probabilistic order in antigenic
 968 variation of *Trypanosoma brucei*. *Int J Parasitol*. 2005;35: 961–972.
- 969 32. Lythgoe KA, Morrison LJ, Read AF, Barry JD. Parasite-intrinsic factors can
 970 explain ordered progression of trypanosome antigenic variation. *Proc Natl Acad*
 971 *Sci U S A*. 2007;104: 8095–8100.
- 972 33. Silva Pereira S, Heap J, Jones AR, Jackson AP. VAPPER: High-throughput
 973 variant antigen profiling in African trypanosomes of livestock. *Gigascience*.
 974 2019;8. doi:10.1093/gigascience/giz091
- 975 34. Pereira SS, Casas-Sánchez A, Haines LR, Ogugo M, Absolomon K, Sanders M,
 976 et al. Variant antigen repertoires in *Trypanosoma congolense* populations and
 977 experimental infections can be profiled from deep sequence data using universal
 978 protein motifs. *Genome Res*. 2018;28: 1383–1394. doi:10.1101/gr.234146.118
- 979 35. Silva Pereira S, de Almeida Castilho Neto KJG, Duffy CW, Richards P, Noyes H,
 980 Ogugo M, et al. Variant antigen diversity in *Trypanosoma vivax* is not driven by
 981 recombination. *Nat Commun*. 2020;11: 844. doi:10.1038/s41467-020-14575-8
- 982 36. Hutchinson OC, Picozzi K, Jones NG, Mott H, Sharma R, Welburn SC, et al.
 983 Variant Surface Glycoprotein gene repertoires in *Trypanosoma brucei* have
 984 diverged to become strain-specific. *BMC Genomics*. 2007;8: 234.
 985 doi:10.1186/1471-2164-8-234
- 986 37. Ilboudo H, Noyes H, Mulindwa J, Kimuda MP, Koffi M, Kaboré JW, et al.
 987 Introducing the TrypanoGEN biobank: A valuable resource for the elimination of
 988 human African trypanosomiasis. *PLoS Negl Trop Dis*. 2017;11: e0005438.
 989 doi:10.1371/JOURNAL.PNTD.0005438
- 990 38. Camara M, Camara O, Ilboudo H, Sakande H, Kaboré J, N'Dri L, et al. Sleeping
 991 sickness diagnosis: use of buffy coats improves the sensitivity of the mini anion
 992 exchange centrifugation test. *Trop Med Int Heal*. 2010;15: 796–799.
 993 doi:10.1111/J.1365-3156.2010.02546.X
- 994 39. Simarro PP, Cecchi G, Paone M, Franco JR, Diarra A, Ruiz JA, et al. The Atlas of
 995 human African trypanosomiasis: A contribution to global mapping of neglected
 996 tropical diseases. *Int J Health Geogr*. 2010;9: 57. doi:10.1186/1476-072X-9-57
- 997 40. González-Andrade P, Camara M, Ilboudo H, Bucheton B, Jamonneau V,
 998 Deborggraeve S. Diagnosis of trypanosomatid infections: Targeting the spliced
 999 leader RNA. *J Mol Diagnostics*. 2014. doi:10.1016/j.jmoldx.2014.02.006
- 1000 41. Brakenhoff RH, Schoenmakers JG, Lubsen NH. Chimeric cDNA clones: a novel
 1001 PCR artifact. *Nucleic Acids Res*. 1991;19: 1949. doi:10.1093/NAR/19.8.1949
- 1002 42. Meyerhans A, Vartanian J-P, Wain-Hobson S. DNA recombination during PCR.
 1003 *Nucleic Acids Res*. 1990;18: 1687–1691. doi:10.1093/NAR/18.7.1687
- 1004 43. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, et al. Full-
 1005 length transcriptome assembly from RNA-Seq data without a reference genome.

- Nat Biotechnol. 2011;29: 644–652.
44. Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*. 2006;22: 1658–1659.
45. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol*. 2009;10: R25.
46. Storrval H, Ramsköld D, Sandberg R. Efficient and comprehensive representation of uniqueness for next-generation sequencing by minimum unique length analyses. *PLoS One*. 2013;8: e53822.
47. Conway JR, Lex A, Gehlenborg N. UpSetR: An R Package for the Visualization of Intersecting Sets and their Properties. doi:10.1101/120600
48. Wickstead B, Gull K. Dyneins across eukaryotes: a comparative genomic analysis. *Traffic*. 2007;8: 1708–1721.
49. Howe K, Bateman A, Durbin R. QuickTree: Building huge neighbour-joining trees of protein sequences. *Bioinformatics*. 2002;18: 1546–1547. doi:10.1093/bioinformatics/18.11.1546
50. Paradis E, Schliep K. Ape 5.0: An environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics*. 2019;35: 526–528. doi:10.1093/bioinformatics/bty633
51. Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol Syst Biol*. 2011;7: 539–539. doi:10.1038/msb.2011.75
52. Eddy SR. Accelerated Profile HMM Searches. *PLOS Comput Biol*. 2011;7: e1002195. doi:10.1371/JOURNAL.PCBI.1002195
53. Csárdi G, Nepusz T. The igraph software package for complex network research.
54. Newman MEJ. Finding community structure in networks using the eigenvectors of matrices. *Phys Rev E*. 2006;74: 036104. doi:10.1103/PhysRevE.74.036104
55. Lawrence M, Huber W, Pagès H, Aboyoun P, Carlson M, Gentleman R, et al. Software for Computing and Annotating Genomic Ranges. *PLOS Comput Biol*. 2013;9: e1003118. doi:10.1371/JOURNAL.PCBI.1003118
56. González-Andrade P, Camara M, Ilboudo H, Bucheton B, Jamonneau V, Deborggraeve S. Diagnosis of trypanosomatid infections: targeting the spliced leader RNA. *J Mol Diagn*. 2014;16: 400–404.
57. Berberof M, Pérez-Morga D, Pays E. A receptor-like flagellar pocket glycoprotein specific to *Trypanosoma brucei gambiense*. *Mol Biochem Parasitol*. 2001;113: 127–138.
58. Weirather JL, Wilson ME, Donelson JE. Mapping of VSG similarities in *Trypanosoma brucei*. *Mol Biochem Parasitol*. 2012;181: 141–152.
59. Mulindwa J, Leiss K, Ibberson D, Kamanyi Marucha K, Helbig C, Melo do Nascimento L, et al. Transcriptomes of *Trypanosoma brucei rhodesiense* from sleeping sickness patients, rodents and culture: Effects of strain, growth conditions and RNA preparation methods. Salavati R, editor. *PLoS Negl Trop Dis*. 2018;12: e0006280.
60. De Greef C, Hamers R. The serum resistance-associated (SRA) gene of *Trypanosoma brucei rhodesiense* encodes a variant surface glycoprotein-like protein. *Mol Biochem Parasitol*. 1994;68: 277–284.

61. Jackson AP, Sanders M, Berry A, McQuillan J, Aslett MA, Quail MA, et al. The genome sequence of *Trypanosoma brucei gambiense*, causative agent of chronic human African Trypanosomiasis. *PLoS Negl Trop Dis*. 2010;4. doi:10.1371/journal.pntd.0000658
62. Siström M, Evans B, Benoit J, Balmer O, Aksoy S, Caccone A. De Novo Genome Assembly Shows Genome Wide Similarity between *Trypanosoma brucei brucei* and *Trypanosoma brucei rhodesiense*. Louis EJ, editor. *PLoS One*. 2016;11: e0147660. doi:10.1371/journal.pone.0147660
63. Siström M, Evans B, Björnson R, Gibson W, Balmer O, Maser P, et al. Comparative genomics reveals multiple genetic backgrounds of human pathogenicity in the *trypanosoma brucei* complex. *Genome Biol Evol*. 2014;6: 2811–2819. doi:10.1093/gbe/evu222
64. Mugnier MR, Stebbins CE, Papavasiliou FN. Masters of Disguise: Antigenic Variation and the VSG Coat in *Trypanosoma brucei*. *PLOS Pathog*. 2016;12: e1005784. doi:10.1371/journal.ppat.1005784
65. Aitcheson N, Talbot S, Shapiro J, Hughes K, Adkin C, Butt T, et al. VSG switching in *Trypanosoma brucei*: antigenic variation analysed using RNAi in the absence of immune selection. *Mol Microbiol*. 2006;57: 1608–1622. doi:10.1111/j.1365-2958.2005.04795.x.VSG
66. Liu D, Albergante L, Newman TJ, Horn D. Faster growth with shorter antigens can explain a VSG hierarchy during African trypanosome infections: a feint attack by parasites. *Sci Rep*. 2018;8: 10922. doi:10.1038/s41598-018-29296-8
67. Chappuis F, Loutan L, Simarro P, Lejon V, Büscher P. Options for field diagnosis of human african trypanosomiasis. *Clin Microbiol Rev*. 2005;18: 133–146.
68. Truc P, Lejon V, Magnus E, Jamonneau V, Nangouma A, Verloo D, et al. Evaluation of the micro-CATT, CATT/*Trypanosoma brucei gambiense*, and LATEX/T b *gambiense* methods for serodiagnosis and surveillance of human African trypanosomiasis in West and Central Africa. *Bull World Health Organ*. 2002;80: 882–886.
69. Tonkin-Hill G, Ruybal-Pesántez S, Tiedje KE, Rougeron V, Duffy MF, Zakeri S, et al. Evolutionary analyses of the major variant surface antigen-encoding genes reveal population structure of *Plasmodium falciparum* within and between continents. *PLOS Genet*. 2021;17: e1009269. doi:10.1371/JOURNAL.PGEN.1009269
70. Büscher P, Bart JM, Boelaert M, Bucheton B, Cecchi G, Chitnis N, et al. Do Cryptic Reservoirs Threaten Gambiense-Sleeping Sickness Elimination? *Trends Parasitol*. 2018;34: 197–207. doi:10.1016/J.PT.2017.11.008
71. Kirchgatter K, del Portillo HA. Association of Severe Noncerebral *Plasmodium falciparum* Malaria in Brazil With Expressed PfEMP1 DBL1 α Sequences Lacking Cysteine Residues. *Mol Med* 2002 81. 2002;8: 16–23. doi:10.1007/BF03401999
72. Bull PC, Berriman M, Kyes S, Quail MA, Hall N, Kortok MM, et al. *Plasmodium falciparum* Variant Surface Antigen Expression Patterns during Malaria. *PLOS Pathog*. 2005;1: e26. doi:10.1371/JOURNAL.PPAT.0010026
73. Abdi AI, Hodgson SH, Muthui MK, Kivisi CA, Kamuyu G, Kimani D, et al. *Plasmodium falciparum* malaria parasite var gene expression is modified by host antibodies: longitudinal evidence from controlled infections of Kenyan adults with

1098 varying natural exposure. BMC Infect Dis 2017 171. 2017;17: 1–11.
1099 doi:10.1186/S12879-017-2686-0
1100 74. Kyriacou HM, Stone GN, Challis RJ, Raza A, Lyke KE, Thera MA, et al.
1101 Differential var gene transcription in Plasmodium falciparum isolates from patients
1102 with cerebral malaria compared to hyperparasitaemia. Mol Biochem Parasitol.
1103 2006;150: 211–218. doi:10.1016/J.MOLBIOPARA.2006.08.005
1104 75. Duffy F, Bernabeu M, Babar PH, Kessler A, Wang CW, Vaz M, et al. Meta-
1105 analysis of plasmodium falciparum var signatures contributing to severe Malaria in
1106 African children and Indian adults. MBio. 2019;10. doi:10.1128/MBIO.00217-19
1107 76. Pereira SS, Casas-Sánchez A, Haines LR, Ogugo M, Absolomon K, Sanders M,
1108 et al. Variant antigen repertoires in Trypanosoma congolense populations and
1109 experimental infections can be profiled from deep sequence data using universal
1110 protein motifs. Genome Res. 2018;28: 1383. doi:10.1101/GR.234146.118
1111 77. Silva Pereira S, Jackson AP, Figueiredo LM. Evolution of the variant surface
1112 glycoprotein family in African trypanosomes. Trends Parasitol. 2021.
1113 doi:10.1016/J.PT.2021.07.012
1114 78. Pinger J, Nešić D, Ali L, Aresta-Branco F, Lilic M, Chowdhury S, et al. African
1115 trypanosomes evade immune clearance by O-glycosylation of the VSG surface
1116 coat. Nat Microbiol. 2018;19: 53.
1117
1118